COOLEY LLP
BOBBY GHAJAR (198719)
(bghajar@cooley.com)
COLETTE GHAZARIAN (322235)
(cghazarian@cooley.com)
1333 2nd Street, Suite 400
Santa Monica, California 90401
Telephone: (310) 883-6400

MARK WEINSTEIN (193043)
(mweinstein@cooley.com)
KATHLEEN HARTNETT (314267)
(khartnett@cooley.com)
JUDD LAUTER (290945)
(jlauter@cooley.com)
ELIZABETH L. STAMESHKIN (260865)
(lstameshkin@cooley.com)
3175 Hanover Street
Palo Alto, CA 94304-1130
Telephone: (650) 843-5000

CLEARY GOTTLIEB STEEN & HAMILTON LLP
ANGELA L. DUNNING (212047)
(adunning@cgsh.com)
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304
Telephone: (650) 815-4131

PAUL, WEISS, RIFKIN, WHARTON & GARRISON LLP
KANNON K. SHANMUGAM (*pro hac vice*)
(kshanmugam@paulweiss.com)
2001 K Street, NW
Washington, DC 20006
Telephone: (202) 223-7300

Counsel for Defendant Meta Platforms, Inc.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

RICHARD KADREY, *et al.*, Individual and Representative Plaintiffs, v. META PLATFORMS, INC., a Delaware corporation; Defendant.	Case No. 3:23-cv-03417-VC-TSH **DECLARATION OF DAVID ESIOBU IN SUPPORT OF META'S MOTION FOR PARTIAL SUMMARY JUDGMENT**

I, David Esiobu, declare:

1. I am over the age of 18 and am competent to make this declaration. I am a Software Engineer in the Generative AI ("Gen AI") division of Meta Platforms, Inc. ("Meta"). I have been employed by Meta since August 2021. I have personal knowledge of the facts contained in this declaration in support of Defendant Meta Platform Inc.'s Motion for Partial Summary Judgment. I declare that the following is true to the best of my knowledge, information, and belief, and that if called upon to testify, I could and would testify to the following.

Professional Background

2. In 2007, I received a Bachelor's degree in Computer Engineering from Georgia Institute of Technology. I have worked as a software engineer from January 2008 to the present, including at Citrix, Amazon, Microsoft, and Meta.

"Memorization" Analyses and Associated Mitigations

3. As part of my work at Meta, I worked on analyses designed to measure the rate of "memorization" of training data in Llama models, that is, whether the Llama models were capable of reproducing significant portions of their training data verbatim, and if so, to measure the rate of such reproduction. One reason Meta wanted to analyze whether the Llama models were reproducing training data was to determine their likelihood of reproducing copyrighted material, which Meta wanted to avoid.

4. Generally speaking, the memorization analyses with which I was involved included the steps of (a) prompt sampling, (b) response sampling, and (c) scoring. Prompt sampling is a process to create a test set of the training data, for example, by selecting from the training data 10,000 spans of 50 tokens. Next, in response sampling, we provide a predetermined number of tokens of training data as prompts to the model, and collect responses from the model. Our analysis used a technique known as "greedy sampling," in which the model generates responses by selecting

the token with the highest probability at each step to maximize the chances that it would reproduce the training text and thereby conservatively estimate the likelihood of memorization. In the scoring step, we determine if the responses output by the model matched at least the next 50 tokens found in the original training data, to determine how many tokens (if any) the model was able to correctly reproduce. I note that determining whether the model can accurately reproduce 50 tokens from the original training data (roughly corresponding to 25-30 words or 2-3 sentences) is a commonly used technique and benchmark for studying memorization in large language models.

5. We performed these memorization analyses across a number of different categories of text data, including random samples from books in the "Books3" dataset based on similarity to titles from The New York Times's Top-100 Bestseller's list. We performed tests on both pretrained versions of Llama 2 and 3 without fine-tuning, and on fine-tuned versions of Llama 2 and 3. The rate of memorization for the Llama 2 and 3 models that we tested was generally quite low, and for the fine-tuned versions, even lower (in some cases, near zero).

6. I am also aware of certain mitigations that were implemented at Meta to further reduce the risk of reproduction of training data, which includes deduplication of data, and reducing the number of times the models trained on certain data (referred to as the number of "epochs"). The theory behind these mitigations is that excessive repeated training on certain data may increase the likelihood that the model is able to reproduce that data as output.

AWS Instance Set-Up

7. In March of 2024, I set up six Amazon Web Services ("AWS") instances for the team to use to download additional books related data from a website called Anna's Archive. These instances used our default configuration for AWS instances, which was designed to block inbound connection requests unless Meta was the originator of the connection, except for specific types of internal Meta requests.

I declare under penalty of perjury that the foregoing is true and correct. Executed on this 24th day of March, at Seattle, WA.

/s/
David Esiobu